Results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition

Leon Derczynski
University of Sheffield
S1 4DP, UK
leon.d@shef.ac.uk

Eric Nichols
Honda Research Institute
Saitama, Japan
e.nichols@jp.honda-ri.com

Marieke van Erp
VU University Amsterdam
Amsterdam, Netherlands
marieke.van.erp@vu.nl

Nut Limsopatham
Accenture
Dublin, Ireland
nut.limsopatham@gmail.com

Abstract

This shared task focuses on identifying unusual, previously-unseen entities in the context of emerging discussions. Named entities form the basis of many modern approaches to other tasks (like event clustering and summarization), but recall on them is a real problem in noisy text - even among annotators. This drop tends to be due to novel entities and surface forms. Take for example the tweet "so.. kktny in 30 mins?!" – even human experts find the entity *kktny* hard to detect and resolve. The goal of this task is to provide a definition of emerging and of rare entities, and based on that, also datasets for detecting these entities. The task as described in this paper evaluated the ability of participating entries to detect and classify novel and emerging named entities in noisy text.

1 Introduction

Named Entity Recognition (NER) is the task of finding in text special, unique names for specific concepts. For example, in "Going to San Diego", "San Diego" refers to a specific instance of a location; compare with "Going to the city", where the destination isn't named, but rather a generic city.

NER is sometimes described as a solved task due to high reported scores on well-known datasets, but in fact the systems that achieve these scores tend to fail on rarer or previously-unseen entities, making the majority of their performance score up from well-known, well-formed, unsurprising entities (Augenstein et al., 2017). This leaves them ill-equipped to handle NER in new environments (Derczynski et al., 2015). As new named entities are guaranteed to continuously emerge and gradually replace the older ones, it is important to be able to handle this change. This paper gives data and metrics for evaluating the ability of systems to detect and classify novel, emerging, singleton named entities in noisy text, including the results of seven systems participating in the WNUT 2017 shared task on the topic.

One approach to tackle rare and emerging entities would be to continuously create new training data, allow systems to learn the updates and newer surface forms. However, this involves a sustained expense in annotation costs. Another solution is to develop systems that are less sensitive to change, and can handle rare and emerging entity types with ease. This is a route to sustainable NER approaches, pushing systems to generalise well. It is this second approach that the WNUT17 shared task focuses on.

2 Task Definition

With the novel and emerging entities recognition task, we aim to establish a new benchmark dataset and current state-of-the-art for the recognition of entities in the long tail. Most language expressions form a Zipfian distribution (Zipf, 1949; Montemurro, 2001) wherein a small number of very frequent observations occur and a very long tail of less frequent observations. Our research community's benchmark datasets, representing only a small sample of all language expressions, often follow a similar distribution if a standard sample is taken. Recently, an awareness of the limitations of current evaluation datasets has risen (Hovy et al., 2006; van Erp et al., 2016; Postma et al., 2016). Due to this bias and the way many NLP ap-

Proceedings of the 3rd Workshop on Noisy User-generated Text, pages 140–147
Copenhagen, Denmark, September 7, 2017.

proaches work internally (i.e. through deriving a model from the training data that often incorporates frequency information) many NLP systems are predisposed towards the high-frequency observations and less so to low-frequency or unknown observations. This is clearly exhibited in the fact that many NLP systems' scores drop when presented with data that is different in type or distribution from the data it was trained on (Augenstein et al., 2017).

We aim to contribute to mitigating the problem of limited datasets through this shared task, for which we have annotated and made available 2,295 texts taken from three different sources (Reddit, Twitter, YouTube, and StackExchange comments) that focus on entities that are emerging (i.e. not present in data from n years ago) and rare (i.e. not present more than k times in our data).

3 Data

To focus the task on emerging and rare entities, we set out to assemble a dataset where very few surface forms occur in regular training data and very few surface forms occur more than once. Ideally, none of the surface forms would be shared between the training data and test data, but this was too ambitious in the time available.

3.1 Sources and Selection

In this section, we detail the dataset creation.

Training data – Following the WNUT15 task (Baldwin et al., 2015), the dataset from earlier Twitter NER exercises (Ritter et al., 2011) comprised this task's training data. This dataset is made up of 1,000 annotated tweets, totaling 65,124 tokens.

Development and test data – Whilst Twitter is a rich source for noisy user-generated data, we also sought to include texts that were longer than 140 characters as these exhibit different writing styles and characteristics. To align some of the development and test data with the training data, we included Twitter as a source, but additional comments were mined from Reddit, YouTube and StackExchange. These sources were chosen because they are large and samples can be mined along different dimensions such as texts from/about geospecific areas, and about particular topics and events. Furthermore, the terms of use of the sources allowed us to download, store and distribute the data.

Reddit Documents were drawn from comments[1] from various English-speaking subreddits over January-March 2017. These were selected based on volume, for a variety of regions and granularities. For example, country- and city-level subreddits were included, as well as non-geospecific forums like `/r/restaurants`. The full list used was:

Global: politics worldnews news sports soccer restaurants

Anglosphere, low-traffic: bahamas belize Bermuda botswana virginislands Guam isleofman jamaica TrinidadandTobago

Anglosphere, high-traffic: usa unitedkingdom canada ireland newzealand australia southafrica

Cities: cincinnati seattle leeds bristol vancouver calgary cork galway wellington sydney perth johannesburg montegobay

To ensure that comments would be likely to include named entities, the data was pre-empted (Derczynski and Bontcheva, 2015) using proper nouns as an entity-bearing signal. Documents were filtered to include only those between 20 and 400 characters in length, split into sentences, and tagged with the NLTK (Bird, 2006) and Stanford CoreNLP (Manning et al., 2014) (using the GATE English Twitter model (Cunningham et al., 2012; Derczynski et al., 2013)) POS taggers. Only sentences with at least one word that was tagged as `NNP` by both taggers were kept.

YouTube The corpus includes YouTube comments. These are drawn from the all-time top 100 videos across all categories, within certain parts of the anglosphere (specifically the US, the US, Canada, Ireland, New Zealand, Australia, Jamaica, Botswana, South Africa and Singapore) during April 2017. One hundred top-level comments were drawn from each video. Non-English comments were removed with `langid.py` (Lui and Baldwin, 2012). Finally, in an attempt to cut out trite comments and diatribes, comments were filtered for length: min 10, max 200 characters.

Twitter The twitter samples were drawn from time periods matching recent natural disasters, specifically the Rigopiano avalanche and the Palm Sunday shootings. This was intended to select

[1]The raw data can be downloaded from `https://files.pushshift.io/reddit/comments/`.

content about emerging events, that may contain highly-specific and novel toponyms. Content was taken from an archive of the Twitter streaming API, processed to extract English-language documents using `langid.py`, and twokenized (O'Connor et al., 2010).

StackExchange Another set of user-generated contents were drawn from StackExchange[2]. In particular, title posts and comments, which were posted between January-May 2017 and also associated to five topics (including movies, politics, physics, scifi and security) were downloaded from archive.org[3]. From these title posts and comments, 400 samples were uniformly drawn for each topic. Note that title posts and comments that are shorter than 20 characters or longer than 500 characters were excluded, in order to keep the task feasible but still challenging. On average the length of title posts and comments is 118.73 with a standard deviation of 100.89.

Note that the data is of mixed domains, and that the proportions of the mixture are not the same in dev and test data. This is intended to provide a maximally adverse machine learning environment. The underlying goal is to improve NER in a novel and emerging situation, where there is a high degree of drift. This challenges systems to generalise as best they can, instead of e.g. memorising or relying on stable context- or sub-word-level cues. Additionally, we know that entities mentioned vary over time, as does the linguistic context in which entities are situated (Derczynski et al., 2016). Changing the particular variant of noisy, user-generated text somewhat between partitions helps create this environment, high in diversity, and helps represent the constant variation found in the wild.

3.2 Preprocessing

Candidate development and test data was filtered for common entities. To ensure that all entities in the development and test data were novel, surface forms marked as entities in the training data were gathered into a blacklist. Any texts containing any of these surface forms were excluded from the final data.

Texts were tokenized using twokenizer and processed through GATE (Cunningham et al., 2012) for crowdsourcing. The corpus was *not* screened for obscenity and potentially offensive content.

3.3 Data Splits

The development data was taken from YouTube. The test split was drawn from the remaining sources.

3.4 Annotation Guidelines

Various named entity annotation schemes are available for named entity annotation (cf. CoNLL (Sang, 2002), ACE (LDC, 2005), MSM (Rizzo et al., 2016)). Based on these, we annotate the following entity types:

1. `person`
2. `location` (including GPE, facility)
3. `corporation`
4. `product` (tangible goods, or well-defined services)
5. `creative-work` (song, movie, book and so on)
6. `group` (subsuming music band, sports team, and non-corporate organisations)

The following guidelines were used for each class.

`person` – Names of people (e.g. *Virginia Wade*). Don't mark people that don't have their own name. Include punctuation in the middle of names. Fictional people can be included, as long as they're referred to by name (e.g. *Harry Potter*).

`location` – Names that are locations (e.g. *France*). Don't mark locations that don't have their own name. Include punctuation in the middle of names. Fictional locations can be included, as long as they're referred to by name (e.g. *Hogwarts*).

`corporation` – Names of corporations (e.g. *Google*). Don't mark locations that don't have their own name. Include punctuation in the middle of names.

`product` – Name of products (e.g. *iPhone*). Don't mark products that don't have their own name. Include punctuation in the middle of names. Fictional products can be included, as long as they're referred to by name (e.g. *Everlasting Gobstopper*). It's got to be something you can touch, and it's got to be the official name.

[2] https://stackexchange.com

[3] https://archive.org/download/stackexchange

Metric	Dev	Test
Documents	1,008	1,287
Tokens	15,734	23,394
Entities	835	1,040
person	470	414
location	74	139
corporation	34	70
product	114	127
creative-work	104	140
group	39	150

Table 1: The emerging entity dataset statistics

`creative-work` – Names of creative works (e.g. *Bohemian Rhapsody*). Include punctuation in the middle of names. The work should be created by a human, and referred to by its specific name.

`group` – Names of groups (e.g. *Nirvana*, *San Diego Padres*). Don't mark groups that don't have a specific, unique name, or companies (which should be marked `corporation`).

3.5 Annotation

Once selected and preprocessed, annotations were taken from the crowd. The GATE crowdsourcing plugin (Bontcheva et al., 2014) provided effective mediation with CrowdFlower for this. Three annotators were allocated per document/sentence, and all sentences were multiply annotated. Annotators were selected from the UK, USA, Australia, New Zealand, Ireland, Canada, Jamaica and Botswana. Once gathered, crowd annotations were processed using max-recall automatic adjudication, which has proven effective for social media text (Derczynski et al., 2016). The authors performed a final manual annotation over the resulting corpus, to compensate for crowd noise.

3.6 Statistics

The dataset dimensions are given in Table 1. The test partition was slightly larger than the development data, which we hope provides greater resolution on this more critical part.

4 Evaluation

The shared task evaluates against two measures. In addition to classical entity-level precision, recall and their harmonic mean, F1, surface forms found in the emerging entities task are also evaluated. The set of unique surface forms in the gold data and the submission are compared, and their precision, recall and F1 are measured as well. This latter measure measures how good systems are at correctly recognizing a diverse range of entities, rather than just the very frequent surface forms.

For example, the classical measure would reward a system that always recognizes *London* accurately, and so such a system would get a high score on a corpus where 50% of the Location entities are just *London*. The second measure, though, would reward *London* just once, regardless of how many times it appeared in the text.

These two measures are denoted F1 (entity) and F1 (surface).

Surface forms should also be given the right class. For example, finding *London* as an entity is useful, but not if it's recognized as a product. Therefore, when computing surface F1, the units used for evaluation are $\langle surfaceform, entitytype \rangle$ tuples. This favors a certain kind of system construction; for example, the tuple formulation assumes that systems are doing joint recognition and typing, instead of the two in distinct stages. However, our goal is to evaluate performance of systems after both named entity recognition and typing, so it fits well in this use case.

5 Results

Results of the evaluation are given in Table 2. Note that surface recognition performance is often lower than entity recognition performance, suggesting that the entities being missed are those that are rarer, and so don't count towards entity F1 as much. We also see that NER in novel, emerging settings remains hard, reinforcing earlier findings that NE systems do not generalize well, especially in this environment (Augenstein et al., 2017).

6 Analysis

To gain insights into the difficult and less difficult parts of the task, we did a qualitative analysis of the outputs of the different systems. We see the most systems have no problems with entities that consist of common English names (e.g. "Lynda", "Becky"). However, when (part of) a name is also a common word (e.g. "Andrew Little", "Donald Duck"), we see that some systems only identify "Andrew" or "Donald" as part of the name. Furthermore, some systems erroneously tag words such as "swift" as entities, probably due to a bias towards 'Taylor Swift' in many current datasets.

Team	F1 (entity)	F1 (surface)
Arcada (Jansson and Liu, 2017)	39.98	37.77
Drexel-CCI (Williams and Santia, 2017)	26.30	25.26
FLYTXT (Sikdar and Gambäck, 2017)	38.35	36.31
MIC-CIS	37.06	34.25
SJTU-Adapt (Lin et al., 2017)	40.42	37.62
SpinningBytes (von Däniken and Cieliebak, 2017)	40.78	39.33
UH-RiTUAL (Aguilar et al., 2017)	**41.86**	**40.24**

Table 2: Results of the emerging entity extraction task.

Locations that contain elements that are also common in person names present an obstacle for the participating systems, for example in the detection of "Smith Tower" or "Crystal Palace" where "Smith" and "Crystal" are sometimes recognised as person names.

Names originating from other languages such as "Leyonhjelm" or "Zlatan" for persons or "Sonmarg" and "Mahazgund" for locations often present problems for the systems. "Mahagzund" is for example classified as corporation, group or person or "other" (no entity) whilst it refers to a village in Kashmir region of India.

`Corporation` and `creative work` were generally a difficult classes for the systems to predict. For `corporation`, this may be partly due to confusion between the corporation and group and product classes, as well as the fact that sometimes the corporation name is used to indicate a headquarters. For example "Amazon" on its own would in most cases be deemed a corporation in our gold standard, but in "Amazon Web Services" it is part of a product name. The 'White House' can both be a location and a corporation, which requires the systems to distinguish between subtle contextual differences in use of the term.

The difficulty in detecting entities of class `creative-work` can often be explained by the fact that these entities contain person names (e.g. "Grimm") , common words (e.g. "Demolition Man", "Rogue One") and can be quite long (e.g. "Miss Peregrine's Home for Peculiar Children").

Annotation still remains hard; some entities in the corpus, if we co-opt Kripke's "rigid designator" (Kripke, 1972) to define that role, are hard to fit into a single category. There were also other types of entity in the data; we did not attempt to define a comprehensive classification schema. The shortness of texts often makes disambiguation hard, too, as the spatial, temporal, conversational and topical context which a human reader relies on to interpret texts are all hidden under this model of annotation.

Twitter accounts can also fall into a number of different classes, and rather than instruct annotators on this, we left behavior up to them. Much prior work has avoided assigning tags to these (Ritter et al., 2011; Liu et al., 2011) though accounts often represent not only a person, also organizations, regions, buildings and so on. Therefore, much of our data carries these labels on Twitter account names, where the annotator has specified it.

7 Related Work

Named entity recognition has a long standing tradition of shared tasks, with the most prominent being the multilingual named entity recognition tasks organised at CoNLL in 2002 and 2003 (Sang, 2002; Tjong Kim Sang and Meulder, 2003). However, these, as well as follow-up tasks such as ACE (LDC, 2005) focused on formal and relatively clean texts such as newswire. This remains a difficult task, especially with the addition of the OntoNotes dataset, with modern work still pushing forward the state of the art (Chiu and Nichols, 2016).

Since 2011, Twitter has been gaining attention as a rich source for information extraction challenges such as (Ritter et al., 2011) and the Making Sense of Microposts challenge series starting in 2013 (Rizzo et al., 2017).

Emerging entities have received some attention entity linking approaches (Hoffart et al., 2014; far, 2016; NIST, 2017). In particular for entity linking, identifying whether an entity is present in a knowledge base to prevent an erroneous link from being created is a key problem.

Rare entities are an even less researched problem. Recasens et al. (2013) attempt to identify

entity mentions that occur only once within a discourse to improve co-reference resolution. In (Jin et al., 2014), a system is presented that is focused on linking low frequent entities.

In the previous two WNUTs there has been attention for named entity recognition in noisy user-generated data in the form of a shared task on Named Entity Recognition in Twitter (Baldwin et al., 2015; Strauss et al., 2016). However, in those tasks, the dataset consisted of a random sample from a particular period without a particular focus on rare or emerging entities.

8 Conclusion

We have presented the setup and results of the WNUT2017 Shared Task on Novel and Emerging Entity Recognition. For this task, we created a new benchmark dataset consisting of 1,008 development and 1,287 test documents containing nearly 2,000 entity mentions. The documents were chosen in such a way that they contained mostly rare and novel entities of the types `person`, `location`, `corporation`, `product`, `creative-work` and `group`. The results of the seven systems that participated in this task show that entity recognition on these entities indeed is more difficult than on high frequent entities commonly found in named entity recognition challenges. More work in this area is thus needed and this shared task is only a small start. Going forward, datasets like this may be extended, possibly also with other entity classes for particular domains. Furthermore, we hope that more NLP tasks take up the challenge of creating more diverse benchmark datasets to expand our coverage of rare and novel language use.

Finally, the task is very tough. These are low figures for named entity recognition, and the surface form capture was even harder, reinforcing earlier findings that systems are failing to generalise successfully, instead profiting from frequently repeated entities in regular contexts. This is not working for noisy text, not Tweets, but broadly.

Acknowledgments

We thanks the participants for their enjoyable collaboration and for joining in this new task. This research received support from the European Commission's Horizon 2020 funding programme under grant agreement 687847, COMRADES. Leon Derczynski thanks the University of California San Diego for facilities provided during this research. Marieke van Erp acknowledges that the research for this paper was made possible by the CLARIAH-CORE project financed by NWO.

References

2016. *On Emerging Entity Detection*.

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A Multi-task Approach for Named Entity Recognition in Social Media Data. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. 2017. Generalisation in Named Entity Recognition: A Quantitative Analysis. *Computer Speech & Language* .

Timothy Baldwin, Young-Bum Kim, Marie Catherine De Marneffe, Alan Ritter, Bo Han, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *ACL-IJCNLP* 126:2015.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*. Association for Computational Linguistics, pages 69–72.

Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Dominic Paul Rout. 2014. The GATE Crowdsourcing Plugin: Crowdsourcing Annotated Corpora Made Easy. In *EACL*. pages 97–100.

Jason PC Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics* 4:357–370.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus Roberts, Danica Damljanovic, et al. 2012. Developing language processing components with gate version 8 (a user guide). *University of Sheffield, UK, Web: http://gate.ac.uk/sale/tao/index.html* .

Leon Derczynski and Kalina Bontcheva. 2015. Efficient named entity annotation through pre-empting. In *International Conference Recent Advances in Natural Language Processing, RANLP*. Association for Computational Linguistics, volume 2015, pages 123–130.

Leon Derczynski, Kalina Bontcheva, and Ian Roberts. 2016. Broad Twitter Corpus: A Diverse Named Entity Recognition Resource. In *In Proc. of the Intl Conference on Computational Linguistics (COLING)*. pages 161–172.

Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, Johann Petrak, and Kalina Bontcheva. 2015. Analysis of named entity recognition and linking for tweets. *Information Processing & Management* 51(2):32–49.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *RANLP*. pages 198–206.

Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. 2014. Discovering emerging entities with ambiguous names. In *Proceedings of the 23rd international conference on World wide web*. pages 385–396.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. Association for Computational Linguistics, New York City, USA, pages 57–60.

Patrick Jansson and Shuhua Liu. 2017. Distributed Representation, LDA Topic Modelling and Deep Learning for Emerging Named Entity Recognition from Social Media. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

Yuzhe Jin, Emre Kcman, Kuansan Wang, and Ricky Loynd. 2014. Entity linking at the tail: sparse signals, unknown entities, and phrase models. In *Proceedings of the 7th ACM international conference on Web search and data mining*. pages 453–462.

Saul A Kripke. 1972. Naming and necessity. In *Semantics of natural language*, Springer, pages 253–355.

LDC. 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Entities version 5.6.1*. Linguistic Data Consortium.

Bill Y. Lin, Frank Xu, Zhiyi Luo, and Kenny Zhu. 2017. Multi-channel BiLSTM-CRF Model for Emerging Named Entity Recognition in Social Media. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. 2011. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, pages 359–367.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, pages 25–30.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL (System Demonstrations)*. pages 55–60.

Marcelo A Montemurro. 2001. Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications* 300(3):567–578.

NIST. 2017. Tac kbp2017 entity discovery and linking pilot on 10 low-resource languages. Technical report, NIST.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proceedings of ICWSM-2010 (demo track)*.

Marten Postma, Filip Ilievski, Piek Vossen, and Marieke van Erp. 2016. Moving away from semantic overfitting in disambiguation datasets. In *Proceedings of EMNLP 2016's UBLP (Uphill Battles in Language Processing) workshop*.

Marta Recasens, Marie-Catherine de Marneffe, and Christopher Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *HLT-NAACL*. pages 627–633.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*. Edinburgh, UK.

Giuseppe Rizzo, Bianca Pereira, Andrea Varga, Marieke van Erp, and Amparo Elizabeth Cano Basave. 2017. Lessons learnt from the Named Entity rEcognition and Linking (NEEL) challenge series. *Semantic Web Journal* .

Giuseppe Rizzo, Marieke van Erp, Julien Plu, and Raphaël Troncy. 2016. Making sense of microposts (#microposts2016) named entity recognition and linking (neel) challenge. In *Proceedings of the 6th Workshop on 'Making Sense of Microposts' co-located with the 25th International World Wide Web Conference (WWW 2016)*.

Erik F. Tjong Kim Sang. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-2002*. Taipei, Taiwan.

Utpal Kumar Sikdar and Björn Gambäck. 2017. A Feature-based Ensemble Approach to Recognition of Emerging and Rare Named Entities. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

Benjamin Strauss, Bethany E Toma, Alan Ritter, Marie-Catherine de Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. *WNUT 2016* page 138.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of CoNLL-2003*. Edmonton, Canada, pages 142–147.

Marieke van Erp, Pablo Mendes, Heiko Paulheim, Filip Ilievski, Julien Plu, Giuseppe Rizzo, and Joerg Waitelonis. 2016. Evaluating entity linking: An analysis of current benchmark datasets and a roadmap for doing a better job. In *Proceedings of LREC 2016*.

Pius von Däniken and Mark Cieliebak. 2017. Transfer Learning and Sentence Level Features for Named Entity Recognition on Tweets . In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

Jake Williams and Giovanni Santia. 2017. Context-Sensitive Recognition for Emerging and Rare Entities. In *Proceedings of the 3rd Workshop on Noisy, User-generated Text (W-NUT) at EMNLP*. ACL.

George Kingsley Zipf. 1949. *Human behavior and the principle of least effort*. Addison-Wesley Press.